A magyar Wikipédia automatikus bejárása és elemzése

Simkó Marcell[1], Góth Júlia[2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter u. 50/a, Magyarország
simko.marcell@hallgato.ppke.hu
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter u. 50/a, Magyarország
goth.julia@itk.ppke.hu

Kivonat: A mai tudományos élet és hétköznapok elengedhetetlen segédeszköze a Wikipédia. Az ingyenes adatbázis hatalmas mennyiségű információt tesz elérhetővé bárki számára. Az információ azonban nem csak az egyes szócikkek szövegében van, hanem a cikkek összekapcsoltságában is, melyet az oldalakon található hiperhivatkozások adnak. Ebben a cikkben feltérképezzük a magyar Wikipédia gráfszerkezetét, majd elemzéseket és vizualizációkat végzünk. Megállapítjuk, hogy a Wikipédia leírására használható a „kis világ" jelenség. A szócikkek szöveges tartalmát is felhasználva megvizsgáljuk a szavak hatványtörvény szerinti eloszlását, és ezt összevetjük a Zipf törvénnyel.

1. Bevezetés

A tényszerű információk visszakeresésének, lekérdezésének egyik legfontosabb és legtöbbet használt eszköze a mai világban kétségkívül a Wikipédia. Emiatt vált érdekessé a Wikipédia hálózatként való szemlélete és a benne rejlő struktúra feltérképezése. A több száz nyelven létező, sok millió lapot tartalmazó weblapon jelen lévő adatmennyiség hatalmas méretének köszönhetően lehetségessé válik olyan kutatások elvégzése, melyek a felhasználók által összegyűjtött tudásban rejlő struktúrát térképezik fel. A nagy mennyiségű természetes nyelvű szöveg elemzésével pedig kvantitatív mérőszámok alapján összehasonlítást tudunk végezni különböző nyelvek (magyar és angol) között.

Az adathalmaz kiválasztásakor figyelembe vettük, hogy nagy, de még kezelhető mennyiségű szöveges adatra van szükség, melyben az információ több, jól elkülönülő témára van osztva. Ilyen szempontból a Wikipédia ideális, hiszen az egyes lapok eleve egy tematikus bontást jelentenek, a hiperszövegben lévő linkek pedig a lapok között felállítanak egy gráf szerű struktúrát. A webhely ingyenes és szabadon hozzáférhető volta, valamint uniform stílusa pedig könnyűvé teszi annak automatikus vizsgálatát. A többi nyelvhez képest a magyar nyelvű Wikipédia közepes méretűnek mondható, melynek köszönhetően a feldolgozás viszonylag gyorsan elvégezhető, viszont mégis rendelkezésünkre áll elég adat megbízható statisztikák elkészítéséhez.

A magyar Wikipédiát mint hálózatot két megközelítésből is vizsgáljuk: a Wikipédia oldalakon lévő linkstruktúra alapján, valamint a Wikipédia oldalak szócikkeinek szö-

veges tartalma alapján. Egy speciálisan erre a célra írt keresőrobot a Wikipédia kezdőlapjától indulva bejárta az összes magyar lapot, és elmentette a linkek által kifeszített gráfot. Mivel csak a valódi tartalommal rendelkező lapok érdekeltek minket, a meta jellegű, illetve az adott lapon belülre mutató linkeket (pl. vitalap, szerkesztési lap stb.) figyelmen kívül hagytuk, a Wikipédián kívülre mutató hivatkozásokkal együtt. A végeredményként kapott gráf több mint 300 ezer csúccsal és 2 millió éllel rendelkezik, mely kifejezetten nehézzé teszi a gráf vizualizációját. Különböző statisztikai jellemzők alapján kiválasztunk részgráfokat, és szemléletesen ábrázoljuk őket.

Megvizsgáljuk, melyek a számok szerint legfontosabbnak tűnő szócikkek, olyan jellemzők alapján, mint pl. legtöbbet hivatkozott lap. A teljes gráfon kvantitatív leírókat (klaszterezési együttható/klikkesedés, átlagos legrövidebb távolság) számolunk ki. Ezen leírók segítségével megállapítjuk, hogy mivel a Wikipédia gráfra nagy klikkesedés, és kis legrövidebb távolság jellemző, leírására használhatjuk a „kis világ" jelenséget, melyet már sok, valós életben előforduló gráfon megfigyeltek, az élet legkülönfélébb területein (fehérjehálózatok, telefonhívások, baráti hálózatok stb.).

A hiperlinkek által leírt gráf struktúra mellett vizsgálatokat végeztünk a szócikkek szövegén is. A szöveg szavakra bontása, majd szótövesítés elvégzése után megszámoljuk a szógyakoriságokat, és lemérjük a hatványtörvény szerinti eloszlás paramétereit. Ezt összevetjük a Zipf törvénnyel. Magyar nyelvű szöveges korpuszokon már végeztek vizsgálatokat korábban is, ami egy adott szűkebb terület szöveges dokumentumait vizsgálta, míg az általunk vizsgált Wikipédia oldalak sokkal heterogénebb halmazt képeznek. Az általunk kapott eredményeket összehasonlítjuk korábbi kutatásokkal mind a magyar, mind az angol nyelv tekintetében.

2. Irodalmi áttekintés

A Wikipédia adatbányászati felhasználása, gráf szerkezetének elemzése nem újdonság. Bizonyos kutatások azt a célt szolgálják, hogy új linkek automatikus beszúrásával javítsák az enciklopédiát. Ilyen például [6], mely a szócikkek különböző nyelvű verziói közötti kapcsolatokat vizsgálja, algoritmust készítve a hiányzó linkek automatikus pótlására. Más kutatások [5] a vitalapokon zajló kommunikáció fa szerkezetét vizsgálják a felhasználók közötti interakciók mintázatainak azonosításához, majd ezeket a mintázatokat összehasonlítják a különböző témájú szócikkek esetén. A szerkesztők közötti kapcsolatokat tanulmányozza [1] is, azonban vitalapok helyett az határozza meg a hálózatot, hogy kik szerkesztenek együtt egy lapot. Korrelációt fedeznek fel bizonyos gráfindikátorok és a szócikkek minősége között.

A szócikkek gráfját, valamint a kategóriák gráfját vizsgálja [7]. A kategóriagráf további elemzése során kiderül, hogy az skálafüggetlen, és kis világ tulajdonsággal rendelkezik – ezt a tulajdonságot mi a magyar szócikkek hálózatán mérjük. A cikk ezután felhasználja a kategóriagráfot szemantikus hasonlósági vizsgálatokhoz, természetes nyelvfeldolgozás céljából. A Wikipédia gráf topológiájával foglalkozik [2]. A gráf növekedését szociális hálókhoz hasonlítja, és alkalmazza rá „a gazdag még gazdagabb lesz" szabályt.

A magyar nyelvre specifikusan végeztek kutatást Dominich és társai [3]. Különböző szépirodalmi és webes korpuszokat felhasználva összehasonlítja a magyar szavak gyakoriságának eloszlását a Zipf törvénnyel. Megállapítja továbbá, hogy a magyar nyelv is „kis világ", azonban az általa használt gráf alapja szavak együtt előfordulása volt, mely lényegesen különbözik ennek a cikknek a témájától.

3. A Wikipédia mint gráf

A vizsgálatunk tárgyát képező adathalmaz a Wikipédia gráfszerkezete, mely csúcsok és élek halmazát jelenti. Elméleti szempontból fontos elkülöníteni két feladatot, illetve két gráftípust. Az első feladat az adatok begyűjtése, mely a keresőmotorok botjaihoz (crawler) hasonlóan lapról lapra ugrálva, hiperhivatkozások segítségével történik. A gráf csúcsai az egyes lapok, az irányított élek pedig az egyik lapról a másikra mutató linkek. Mivel konkrétan a Wikipédia képzi az adatgyűjtés tárgyát, a lapok szócikkeket jelentenek, a hivatkozások pedig kizárólag a Wikipédián belülre mutathatnak, egyik szócikkről a másikra.

A másik feladat az adatok elemzése. Ebben a cikkben a vizsgálódás tárgyát képző gráf megegyezik a bejárási gráffal, azonban későbbi kutatásokban célszerű lenne ennél absztraktabb gráfokat vizsgálni, ahol pl. a gráf csúcsai témaköröket jelentenének [7], az összekötő élek pedig valamilyen jelentésbeli kapcsolatokat. Az ilyen vizsgálatok túlmutatnak ennek a cikknek a hatáskörén.

3.1. Adatgyűjtés

Mielőtt a kutatás méréseit el lehetne végezni, először természetesen adatok gyűjtésére van szükség. Ebben a fázisban egy kifejezetten erre a célra írt crawlert használunk, mely a keresőmotorokéhoz hasonlóan jár lapról lapra, azonban velük ellentétben nem törődik azok szöveges tartalmával, egyetlen célja csupán a hivatkozások kigyűjtése.

A crawler vázlatos működése a következő: egy tetszőleges lap (pl. a Kezdőlap) címét berakjuk egy sorba. Amíg ez a sor nem üres, lekérdezzük a következő címhez tartozó HTML kódot, és vesszük a benne található linkeket. Eldobjuk a már látott, vagy haszontalan linkeket (lásd lentebb), majd a maradékot berakjuk a sorba. Nincs szükség prioritás felállítására a soron belül, mert az egész gráfot bejárjuk.

Ha megvizsgáljuk a szócikkekben található hivatkozásokat, rögtön feltűnik, hogy nagy részük számunkra haszontalan, ugyanis:

a) a (magyar) Wikipédián kívülre mutatnak. Ilyenek a hivatkozásjegyzék elemei, a más nyelvű wikire mutató linkek stb.. A helyes linkek szerencsére könnyen felismerhetők, mert úgy kezdődnek, hogy *„/wiki/"*.
b) a Wikipédián belülre mutatnak, de nem szócikkre. Ezek olyan meta jellegű lapokra hivatkoznak, mint pl. vitalap, szerkesztési lap, kategória lapok stb.. Ezen linkek jól meghatározott formátummal rendelkeznek, pl. *„Vita:<link>"*.

c) szócikken belül konkrét fejezetre mutatnak. Az ilyen link nem dobandó el teljesen, de a fejezet információt ki kell vágni az URL-ből. Ezek a linkek *„<szócikk>#<fejezet>”* alakúak.

Az algoritmus sebességének a szűk keresztmetszete a szerverrel való kommunikáció. A Wikipédia szerverek, miután főleg szöveges adatot szolgáltatnak a felhasználóknak, erősen korlátozzák az egy kliensre jutó sávszélességet. A teljes magyar wiki bejárásához kb. 100 órára volt szükség, összesen 19 GB adatot letöltve. A folyamat végeredménye egy 317 ezer csúcsú, 23 millió élű gráf.

3.2. A gráf elemzése

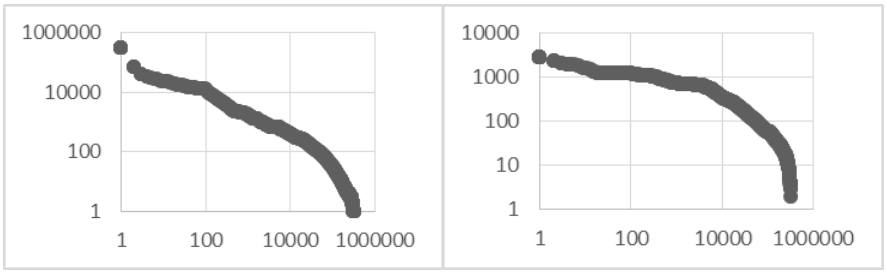
A gráf vizsgálatának a legegyszerűbb és legszemléletesebb módja természetesen a gráf vizualizációja lenne. Sajnos azonban ez a gráf egy nagyságrenddel nagyobb annál, mint amit egy hétköznapi PC-n ábrázolni lehet, ezért érdekessé válik a gráfból releváns részgráfok kiválasztása. Ehhez szükség van valamilyen statisztikai jellemzőre, mellyel kvantitatív módon lehet mérni egy szócikk fontosságát. Ennek kiválasztása azonban közel sem triviális feladat, mint elsőre gondolnánk.

A legegyszerűbb jellemző egyszerűen az adott szócikkre mutató hivatkozások száma, avagy a gráfban a csúcs befoka. Bár az adatgyűjtés során a meta jellegű lapokat már figyelmen kívül hagytuk, a szócikkek között mégis nagyon sok olyan van, melyre nem azért hivatkozik sok szócikk, mert fontosak, hanem egyéb okok miatt. A *Kezdőlap*ra például, – mely formáját tekintve szócikk, de tartalma miatt egészen speciális – minden szócikk hivatkozik. A második leghivatkozottabb lap a *Wikimédia Commons*, a harmadik a *Földrajzi koordináta-rendszer.* Nyilvánvaló, hogy ezek nem a legfontosabb cikkek, csupán azért hivatkozik rájuk sok cikk, mert bizonyos kontextusban mindig relevánsak. (Médiaállományok, illetve földrajzi helyek.) Hasonlóan sokat hivatkozott szócikk csoportok: országok, évszámok, biológiai rendszertannal kapcsolatos cikkek. Az első 10 cikket az 1. táblázatban láthatjuk. Kétféleképpen állíthatjuk fel a sorrendet: hivatkozó egyedi szócikkek száma („Befok”), illetve az összes hivatkozások száma („Befok (többszörös)”). Megnézhetjük természetesen a szócikken található linkeket is, mely a gráfban a kifokszámnak felel meg. A befokhoz hasonlóan itt is pár szócikkfajta dominálja az eredményt, így ez a mérőszám sem jól használható eszköz a lapok fontosságának méréséhez. A legjellemzőbb típusok itt a listák, táblázatok, illetve a sporttal és vasúttal kapcsolatos szócikkek. Ugyanezeket a típusokat látjuk, ha a szócikkek hosszát vesszük alapul. (Nem meglepő módon erős korreláció van egy cikk hossza és a benne lévő linkek száma között.)

Az 1. ábra bemutatja a hivatkozások számának eloszlását az összes szócikken. Csökkenő sorban a 10. szócikktől a 10000.-ig szép hatványtörvény szerinti csökkenést láthatunk, a 10000. cikk után viszont exponenciálisnál is gyorsabb lecsengés jellemző. A hatványtörvény paramétere α=-0.64, illetve α=-0.19 a befok, illetve a kifok esetén. A többszörös hivatkozásszámlálás esetén a grafikonok hasonlóan néznek ki.

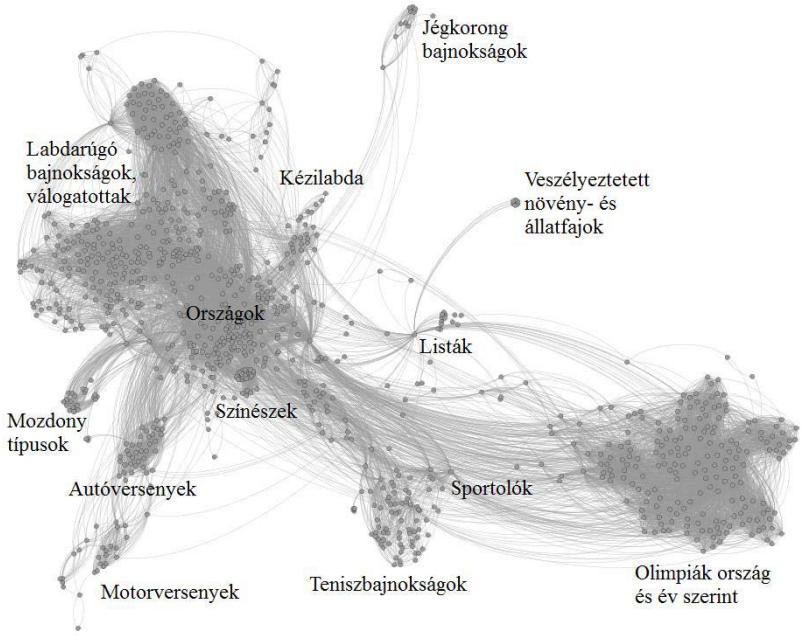
1. táblázat: Szócikkek egyszerű statisztikái. Zárójelben a hivatkozások száma.

	Befok	**Befok (többszörös)**
1.	"Kezdőlap" (316565)	"Kezdőlap" (633239)
2.	"Wikimédia Commons" (74098)	"Wikimédia Commons" (85606)
3.	"Földrajzi koordináta-rendszer" (41227)	"Budapest" (65042)
4.	"Magyarország" (33818)	"Amerikai Egyesült Államok" (59743)
5.	"Amerikai Egyesült Államok" (30819)	"Rend (rendszertan)" (59634)
6.	"Időzóna" (29207)	"Család (rendszertan)" (54555)
7.	"Egyezményes koordinált világidő" (27705)	"Magyarország" (50758)
8.	"Budapest" (27093)	"Osztály (rendszertan)" (44756)
9.	"Rendszertan (biológia)" (24351)	"Törzs (rendszertan)" (42027)
10.	"Ország (rendszertan)" (24186)	"Földrajzi koordináta-rendszer" (41292)
	Kifok	**Kifok (többszörös)**
1.	"Listák listája" (2904)	"Magyar névnapok betűrendben" (9341)
2.	"Magyar névnapok betűrendben" (2467)	"Romániai magyarok listája" (4915)
3.	"Labdarúgó-játékvezetők listája" (2191)	"2014-es labdarúgó-világbajnokság (keretek)" (4452)
4.	"Vegyületek összegképlete" (2043)	"Külföldi festők listája" (3638)
5.	"Vegyületek összegképlet-táblázata" (1955)	"Festőművészek listája" (3626)
6.	"Romániai magyarok listája" (1819)	"2010-es labdarúgó-világbajnokság (keretek)" (3608)
7.	"Katolikus szentek és boldogok listája naptár szerint" (1741)	"Római pápák listája" (3502)
8.	"Labdarúgócsapatok listája" (1656)	"2006-os labdarúgó-világbajnokság (keretek)" (3349)
9.	"A madarak nemeinek listája" (1608)	"Olimpiai érmesek listája atlétikában (férfiak)" (3281)
10.	"2014-es labdarúgó-világbajnokság (keretek)" (1411)	"Katolikus szentek és boldogok listája naptár szerint" (3242)



1. ábra: A befele (balra), illetve kifele (jobbra) irányuló hivatkozások számának eloszlása. Log-log grafikon.

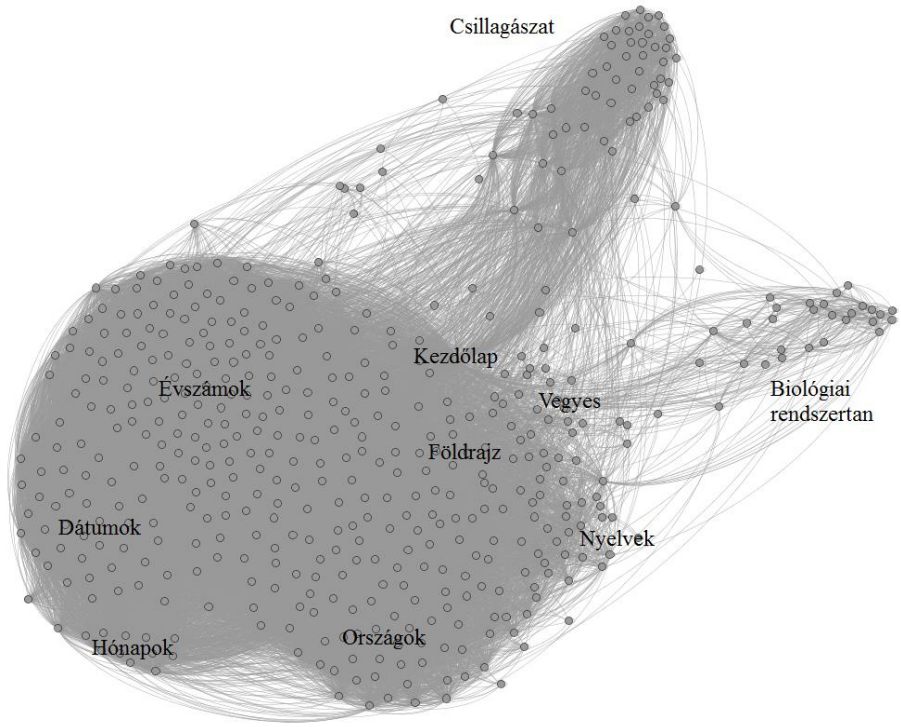
Ezek után ezeket a jellemzőket felhasználhatjuk arra, hogy részgráfokat vizualizáljunk. Bár, amint láthattuk, önmagukban ezek a számok nem tükrözik hűen egy-egy szócikk fontosságát, későbbi kutatások megalapozásaként mégis hasznosak lehetnek.



2. ábra: Az 1000 leghosszabb szócikk.

A 2. ábrán az 1000 leghosszabb szócikket láthatjuk. A gráf megjelenítése a Gephi nevű open-source programmal történt. A programnak semmilyen információja nincs az egyes csúcsokról, pusztán a gráfstruktúrát ismeri. Ennek ellenére a képen jól elkülönülő klaszterekre bomlanak az egyes témakörök. A sportösszefoglalók láthatóan dominálnak, de a lapok szépen elkülönülnek az egyes sportágak szerint. Megfigyelhető, hogy a híres sportolók, a helyszínek, illetve a listák úgynevezett „hub”-ként működnek, nagyon sok kapcsolattal kritikus összekötő láncszemeket alkotva. A 3. ábra az 500 leghivatkozottabb szócikket szemlélteti. Az itt hangsúlyos kategóriák az évszá-

mok és dátumok, de az országok itt is megjelennek. Jól elkülönülve képviseli magát a csillagászat és a biológiai rendszertan.



3. ábra: Az 500 leghivatkozottabb szócikk.

Az eddig tárgyalt felületes leíróknál tovább menve megmértük a gráf klaszterezési (klikkesedési) együtthatóját, valamint a csúcsok közötti átlagos legrövidebb távolságot. Gráfelméletben klikknek nevezzük egy csúcs szomszédságát, ha a szomszédok mind össze vannak egymással kötve. A gyakorlatban a tökéletes nagy klikkek természetesen ritkák, ezért egy aránnyal fejezzük ki, hogy egy adott csúcs szomszédjai mennyire „klikkesednek”. A teljes gráf klaszterezési együtthatója az egyes csúcsok klikkesedéseinek az átlaga. Ezt a mérőszámot összevetjük azzal, amit egy olyan gráfon mérünk, amelynek ugyanennyi csúcsa és éle van, de az élek illeszkedése véletlen. Hasonlóan járunk el az átlagos legrövidebb távolsággal kapcsolatban is. Az eredményeket a 2. táblázat mutatja be. Jól látható, hogy míg a legrövidebb távolság kicsit kisebb, a klaszterezési együttható majdnem 4 nagyságrenddel nagyobb. Ez a két tulajdonság együtt az úgynevezett „kis világ” jelenségre utal. A kis világ számtalan valós életben előforduló gráfban tapasztalható, mint például telefonhívások, szociális hálózatok, fehérje láncok stb.. Fontos megjegyezni, hogy a Wikipédia gráf irányított, ezt mindkét mérőszámnál figyelembe kell venni. Ha a linkek irányítottságát figyelmen kívül hagyjuk, a legrövidebb távolság lecsökkenne 2 alá, a „Kezdőlap” és a „Wikimédia Commons” lapnak köszönhetően. Megjegyzendő, hogy míg [3] is megál-

lapította a Wikipédia kis világ tulajdonságát, az általuk vizsgált gráf egészen más felépítésű.

2. táblázat: Klaszterezési együttható és átlagos legrövidebb távolság.

	Wikipédia	Véletlen gráf
Klaszterezési együttható	23.8%	0.003%
Átlagos legrövidebb távolság	3.92	4.67

4. A Wikipédián belüli szógyakoriságok elemzése

Az előzőekben csak a Wikipédia gráfszerkezetét vizsgáltuk. Értékes információ található azonban a szócikkek szövegében is. Ismert, hogy egy korpuszban a szavak gyakoriságának eloszlása, – mint sok más eloszlás is – hatványtörvényt követ. A hatvány kitevője a Zipf törvény szerint α=-1, azaz egy szó gyakorisága fordított arányosságban áll a sorrendben elfoglalt helyével. Lemértük a magyar Wikipédia szógyakoriságait is.

Korábban az angol Wikipédián végzett kutatás [4] szerint az első 10000 vizsgált szóra igaz a Zipf törvény, azonban az eloszlás végét egy erősebb, α=-2 lecsengés jellemzi. A magyar nyelv tekintetében is történtek mérések, azonban nem a Wikipédiát használva. Dominich és társai [3] különböző szépirodalmi korpuszokat felhasználó kutatása szerint a magyar nyelvre nem igaz a Zipf törvény, mivel a leíró paraméter értéke α=-1.21.

A mérés elvégzéséhez először elő kell készíteni a szöveget. A crawlert módosítottuk, hogy ne a linkeket szedje ki a HTML kódból, hanem a szöveges tartalmat. A szöveget ezután szavakra bontottuk, majd a toldalékok levágásával szótövesítettünk. Összeszámoltuk a szavak előfordulásait, majd csökkenő sorrendet állítottunk fel. Az eloszlást egy log-log grafikonon ábrázoltuk, majd a ráillesztett egyenes meredekségét lemérve megkaptuk az α paramétert.

A mérést elvégezve azt tapasztaltuk, hogy a 10000. szó környékén valóban történik egy törés, ahogy azt [4] is állítja. A magyar nyelv esetében azonban ez a törés közel sem tűnik olyan erősnek, mint az angolban. Az általunk mért értékek: a törés előtt α=-1.36, utána α=-1.66. Eredményeink szerint tehát az eloszlás lényegesen meredekebb, mint azt akár a Zipf törvény, akár [3] állítja.

Hivatkozások

1. Brandes, U., Kenis, P., Lerner, J., van Raaij, D.: Network Analysis of Collaboration Structure in Wikipedia. Proc. of the 18th int. conf. on World wide web (2009) 731–740
2. Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., Caldarelli, G.: Preferential attachment in the growth of social networks: the case of Wikipedia. arXiv:physics/0602026v2 [physics.soc-ph] (2006)
3. Dominich, S., Kiezer T.: Hatványtörvény, „kis világ” és magyar nyelv. Alkalmazott nyelvtudomány, Vol. 5. (2005) 5–24

4. Grishchenko V.: Bouillon project. https://web.archive.org/web/20080217050922/http://oc-co.org/?p=79 (2006)
5. Laniado, D., Tasso, R., Volkovich, Y., Kaltenbrunner, A.: When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. Proc. of the Fifth International AAAI Conf. on Weblogs and Social Media (2011)
6. Sorg, P., Cimiano, P.: Enriching the Crosslingual Link Structure of Wikipedia – A Classification-Based Approach. Proc. of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence (2008)
7. Zesch, T., Gurevych, I.: Analysis of the Wikipedia Category Graph for NLP Applications. Proc. of the TextGraphs-2 Workshop (2007) 1–8